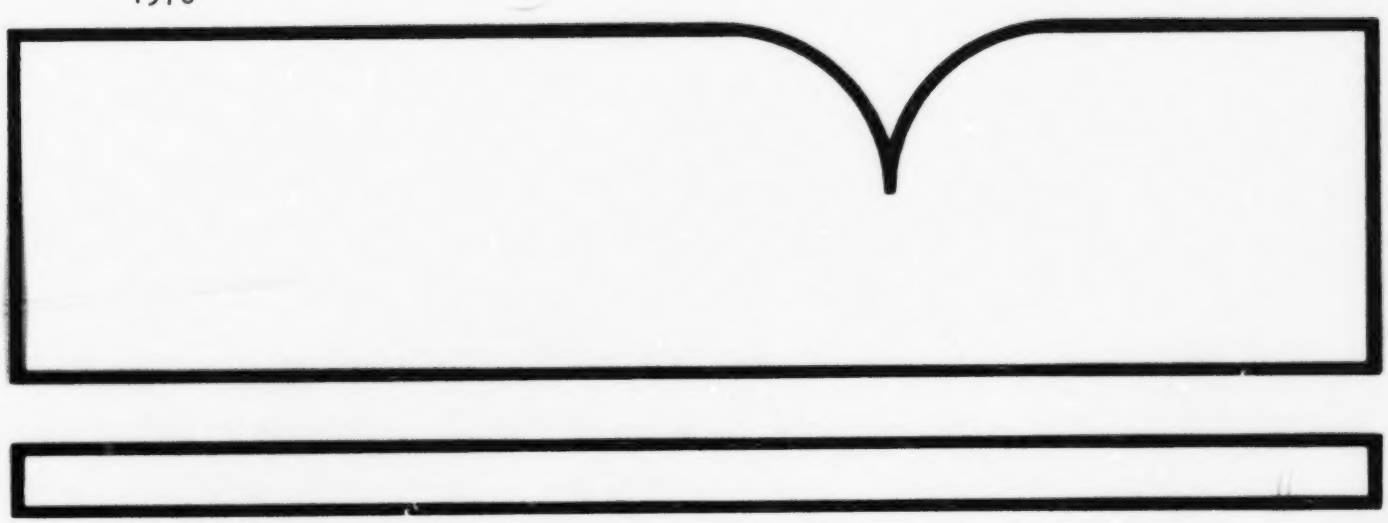
PB80-209471

Refinements of the Lognormal Probability Model for Analysis of Aerometric Data

(U.S.) Environmental Monitoring and Support Lab., Research Triangle Park, NC

1978



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before complet

1. REPORT NO.	2.	3. PB80-209471
EPA-600/J-78-109	JOURNAL ARTICLE	

4. TITLE AND SUBTITLE	5. REPORT DATE
Refinements of the Lognormal Probability Model for Analysis of Aerometric Data	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) David T. Mage, EPA/EMSL/STAB, MD-75, RTP, NC 27711 Wayne R. Ott, EPA/OMTS, RD-680, Washington, DC 20460	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Environmental Monitoring and Support Laboratory Statistical and Technical Analysis Branch MD-75 Research Triangle Park, NC 27711	10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Same	13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA-ORD 600/08

15. SUPPLEMENTARY NOTES
Published in Journal of the Air Pollution Control Association, 28(8):796-798 August 1978

16. ABSTRACT

Logarithmic probability plots of air quality data, commonly described as "approximately straight lines," are shown to have consistent curvature. These curves appear to follow the form of a censored three-parameter lognormal distribution. A procedure for fitting these data is discussed in which a statistical criterion for goodness-of-fit is used to guide the analyst.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field Group
Lognormal distribution Censored distribution Air quality data Statistical tests	probability modeling goodness-of-fit	68A 43F

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
Release Unlimited	20. SECURITY CLASS (This page)	22. PRICE

EPA Form 2220-1 (9-73)

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

Refinements of the Lognormal Probability Model for Analysis of Aerometric Data

David T. Mage and Wayne R. Ott
Office of Research and Development
U. S. Environmental Protection Agency

An important step in the analysis of air quality data is to select an appropriate probability model for the data set under study. Many recent articles have discussed the application of particular probability models to air quality data without considering the comparative fit of different candidate models or the prediction error associated with a particular model.

For example, a series of such articles by Ralph I. Larsen, entitled "An Air Quality Data Analysis System for Interrelating Effects, Standards, and Needed Source Reductions," Parts 1, 2, 3, and 4, have proposed the two parameter[1-3] and three parameter[4] lognormal models for analyzing air quality data in urban areas and near isolated point sources. These models are based on an assumption that "pollutant concentrations are lognormally distributed for all averaging times and thus tend to plot as approximately straight lines on log-probability paper."[4] This assumption is based on an earlier report[5] in which SO_2 concentrations are plotted on logarithmic probability paper giving a line which is "approximately straight" (Figure 3 in Reference 5 and reproduced here as Figure 1). Although useful as a first approximation, many important refinements can be made to this approach.

Dr. Mage is Chief of the Exposure Assessment Branch, Population Studies Division, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711. Dr. Ott is Senior Environ-Engineer, Monitoring Technology Division, Office of Monitoring and Technical Support, U.S. Environmental Protection Agency, Washington, DC 20460.

©Air Pollution Control Association

In 1975, we examined the "straightness" of Larsen's original SO_2 probability plot (Figure 1) and observed that, rather than fitting the standard 2-parameter model, the data described a curve that was better suited to a different distribution, the *censored three-parameter lognormal model* (LN3C).[6,7] This model plots as a curve on logarithmic probability paper. Our error tests with numerous air quality data sets have shown

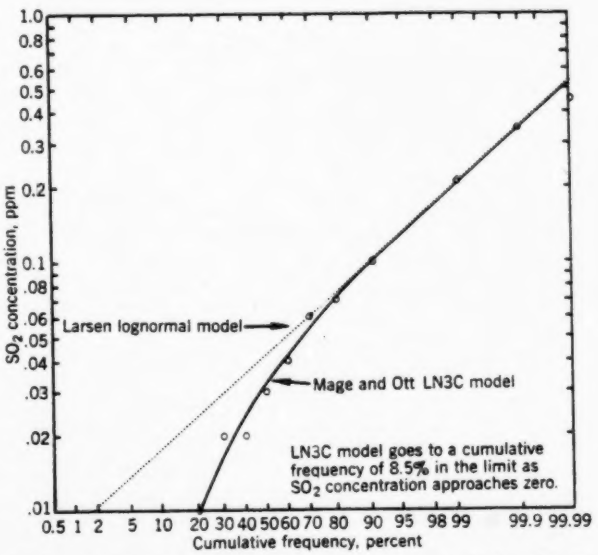


Figure 1. Logarithmic probability plot of 1 hr SO_2 concentration data for Washington, DC, CAMP (Dec 1, 1961–Dec 1, 1968), showing Larsen 2-parameter lognormal model and Mage and Ott censored 3-parameter lognormal model (LN3C).

Reprinted from **APCA JOURNAL,** *Vol. 28, No. 8, August 1978* 796-798

that the LN3C model almost always gives a significant improvement in goodness-of-fit when compared with the standard lognormal model. Although improved fit is obtained whenever a parameter is added to a model, the improvement is sufficiently great that the LN3C model warrants particular consideration by the data analyst.[7]

It should be noted that the term "lognormally distributed," as used by Larsen, is different from its conventional statistical usage. Conventionally, it means that air pollution concentration data arise from some stochastic process in which the logarithm of pollutant concentration is a normally distributed random variable. However, standard statistical goodness-of-fit tests, when applied to air pollution concentration data, usually reject this hypothesis. Papers by Kalpasanov and Kurchatova,[8] who applied the Chi-square test to Bulgarian air quality data, and Curran and Frank,[9] who applied the exponential model to U.S. air quality data, generally question the lognormal assumption. Thus, the lognormal model is supported neither by a physical rationale nor by statistical measures of its validity. Rather, it is intended as a useful practical tool for representing air pollutant concentration frequency distributions.

Because the lognormal model has an engineering basis in these applications rather than a statistical basis, conventional goodness-of-fit tests usually are of reduced significance. Conventional tests usually are designed to test the hypothesis that an original distribution, from which a small random sample is drawn, is not lognormal. For continuous data, the standard tests usually are unsuitable, because the entire original distribution is available (8760 values) and the data are highly serially correlated. Goodness-of-fit measures still are very important in an engineering sense, however, but the concept of "approximate straightness" is not sufficiently well defined in the literature. A small amount of curvature on a logarithmic scale can produce considerable error in predicted values. Furthermore, the same data may give a line which "looks straight" if plotted on Weibull paper or logarithmic probability paper, but even "more straight" on some other probability paper. To address these problems, we are currently developing an engineering goodness-of-fit test and evaluating it with a large number of air quality data sets. We hope to present the findings in a future publication.

Some important technical considerations should be noted in Larsen's manner of applying the two and three parameter models to different data sets. In the two parameter averaging time model, parameters are estimated by the "method of fractiles," fitting the data at the 70% and 99.9% cumulative frequencies. Choosing the two parameters in this fashion causes the arithmetic mean of the model to differ from the arithmetic mean of the data set. Consequently, the convergence point of Larsen's arrowhead diagrams[1-3,5] is not the true arithmetic mean of the data set, and all predictions at various averaging times are displaced accordingly. Because of the basic structure of the averaging time model, it may be preferable to estimate at least one parameter by the "method of moments," in which the mean of the model is set equal to the mean of the data. The second parameter can be estimated from the standard deviation, 70 percentile, 99.9 percentile, or some other percentile of interest.

In Larsen's recent application of the three-parameter lognormal model,[4] this inconsistency is removed, because the Part 4 paper assumes that the arithmetic mean of the model is independent of averaging time and is equal to the measured arithmetic mean. Unfortunately, the paper contains an error, for the examples given do not have this property and conflict with this assumption. In addition, the mean of the three-parameter model is calculated erroneously by allowing pollutant concentrations to be negative. If the negative concentrations predicted by this model are treated as zero, as suggested in the paper, the arithmetic mean of the model will invariably be greater than the mean of the data and will change with averaging time. If the random variable X denotes pollutant concentration, then $X \geq 0$, and the mean μ of the three-parameter lognormal model is properly calculated as follows:

$$\mu = \int_0^\infty x f_X(x)\, dx$$

where $f_X(x)$ = probability density function of the three-parameter lognormal model.

Because Larsen's integral extends to negative values when the third parameter is negative (as it is in the Part 4 paper), this three-parameter model is applied in a manner that is inconsistent with reality. This problem has been solved by introducing censorship to the model, giving the LN3C.[6,7] If readers are aware of this inconsistency in the Part 4 paper, they will avoid confusion when they attempt to apply the 3-parameter model. Whenever the third parameter is negative, the censored model suggested by Mage and Ott[6,7] should be used. With this model, the mean of the data and the mean of the model do not necessarily coincide. Nevertheless, predicted and observed values will match very well.

In the absence of a widely accepted theoretical rationale favoring a specific probability model, the best approach usually is to test candidate models (for example, the lognormal, LN3C, exponential, and Weibull) to determine which model is suited to the problem at hand and the data under study. Goodness-of-fit tests provide important information for this testing process, although they should not be the sole criteria for determining which model is to be selected. Larsen's other two assumptions—that the median concentration and maximum concentration are proportional to averaging time raised to an exponent—are not strictly limited to the lognormal model but can apply to other probability models as well. In the event that no one model is clearly superior, the investigator always should choose the simplest model having the fewest parameters. This may or may not be the lognormal model. It should be noted that the Environmental Protection Agency (EPA) has issued guidelines[10] which call for the data analyst to evaluate carefully the validity of any predictive model he selects. Failure to consider the statistical validity of a model, if hypothesis tests are involved, can lead to predictions which are not supportable when placed under scientific scrutiny.[11]

In conclusion, we believe that the data analyst always should analyze each data set carefully and should not automatically use one particular model. If the analyst, after testing several models, and considering the problem under study, finds that the lognormal model gives the best fit to the data set, then that model should be selected. Results of any goodness-of-fit tests employed in the selection process should accompany the model, and these tests should be tailored to the intended use of the model and the problem being addressed.

References

1. R. I. Larsen, "An air quality data analysis system for interrelating effects, standards, and needed source reductions," *J. Air Poll. Control Assoc.* **23**:993 (1973).
2. R/ I. Larsen, "An air quality data analysis system for interrelating effects, standards, and needed source reductions—Part 2." *J. Air Poll. Control Assoc.* **24**:551 (1974).
3. R. I. Larsen, "An air quality data analysis system for interrelating effects, standards, and needed source reductons—Part 3. Vegetation injury," *J. Air Poll. Control Assoc.* **26**:325 (1976).
4. R. I. Larsen, "An air quality data analysis system for interrelating effects, standards, and needed source reductions—Part 4. A three-parameter averaging-time model," *J. Air Poll. Control Assoc.* **27**:454 (1977).
5. R. I. Larsen, "A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards," Publication No. AP-89, U. S. Environmental Protection Agency, Research Triangle Park, NC, 1971.
6. D. T. Mage and W. R. Ott, "An Improved Statistical Model for Analyzing Air Pollution Concentration Data," Paper No. 75-51.4, 68th Annual Meeting of the Air Pollution Control Association, Boston, MA, 1975.

7. W. R. Ott and D. T. Mage, "A general purpose univariate probability model for environmental data analysis," *Comput. and Ops. Res.* 3;209 (1976).
8. Y. Kalpasanov and G. Kurchatova, "A study of the statistical distribution of chemical pollutants in air," *J. Air Poll. Control Assoc.* 26:981 (1976).
9. T. C. Curran and N. H. Frank, "Assessing the Validity of the Lognormal Model when Predicting Maximum Air Pollutant Concentrations," Paper No. 75-51.3, 68th Annual Meeting of the Air Pollution Control Association, Boston, MA, 1975.
10. "Guidelines for the Interpretation of Air Quality Standards," United States Environmental Protection Agency, Research Triangle Park, NC, OAQPS No. 1.2-008, Revised February 1977.
11. P. E. Wilkins, "Log normal distribution," *J. Air Poll. Control Assoc.* 26:935 (1976).

Comment by Larsen:

Drs. Mage and Ott have correctly noted that the arithmetic mean of the model was not calculated correctly when air quality data were censored at a true concentration of zero in Ref. 4 of their paper. This shortcoming is being corrected in a subsequent paper.

Ralph I. Larsen, Ph.D.
Environmental Research Engineer
Environmental Applications Branch
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
U. S. Environmental Protection Agency